DOCUMENT RESUME

ED 048 373 TM 000 466

AUTHOR McMorris, Robert F.
TITLE Evidence on the Quality of Several Approximations for Commonly Used Measurement Statistics.
PUB DATE Feb 71
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *Data Analysis, Evaluation Techniques, Item Analysis, Measurement Techniques, Predictive Measurement, *Reliability, *Research Methodology, Research Tools, *Standard Error of Measurement, Statistical Analysis, *Statistics, Tests
IDENTIFIERS *Standard Deviation

ABSTRACT

The extent of error likely to occur with each of several approximations for the standard deviation, internal consistency reliability, and the standard error of measurement is analyzed. Approximations were compared with exact statistics obtained on 85 different classroom tests constructed and administered by professors in a variety of fields. Means and standard deviations of the resulting differences supported the use of approximations in practical situations. Results of this analysis suggest a greater number of alternative formulas that might be employed, and provide additional infcrmation concerning the accuracy of approximations with non-normal distributions. (Author)

ED 048373

EVIDENCE ON THE QUALITY OF SEVERAL APPROXIMATIONS FOR COMMONLY USED MEASUREMENT STATISTICS

Robert F. McMorris
SUNY at Albany

What is the extent of error likely with each of several approximations for the standard deviation, internal consistency reliability, and the standard error of measurement? To help answer this question, approximations were compared with exact statistics obtained on 85 different classroom tests constructed and administered by professors in a variety of fields; means and standard deviations of the resulting differences supported the use of approximations in practical situations. Results of this analysis (1) suggest a greater number of alternative formulas that might be employed, and (2) provide additional information concerning the accuracy of approximations with non-normal distributions.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

American Educational Research Association, New York, February 1971

TM 000 466

What is the extent of the error involved with approximations for the standard deviation, internal consistency reliability, and the standard error of measurement? Formulae for approximations were proposed by Jenkins (1946; later unpublished manuscript), Diederich (1958), Saupe (1961), and Lord (1957, 1959), and a modification was proposed by Mason and Odeh (1968). While additional rationales for individual formulae have appeared (Swineford, 1959; Lathrop, 1961; Gardner, 1970), the number of empirical verifications remains limited. Recently Sabers (1970) provided findings supporting the use of short-cut approaches for the standard deviation.

For the present paper approximations for each of the three statistics were computed and compared with the corresponding exact statistics.

METHOD

Data were derived from 85 tests item-analyzed by the Computer Center of the State University of New York at Albany. The classroom tests came from a wide variety of academic disciplines, and naturally differed considerably not only in content but in length and quality. Thus, the tests provided a severe and realistic test of the approximations.

The various formulae for the approximations and the corresponding exact statistics are presented in Table 1. Included in the set of approximations are new ones for reliability formed by substituting approximations in another formula.

The quality of the approximations as estimators of the statistics was described in various ways. The means and standard deviations for the approximations were compared with the same measures for the statistics. The values of the approximations were correlated with those of the statistics, and the relationships shown using scattergrams. Somewhat redundantly, the means and standard deviations for the differences between an

approximation and its statistic were also found.

Indications of conditions which relate to the function of the approximations were gained from relating the difference between approximation and statistic with selected test characteristics: average difficulty, skewness, number of testees, and number of items. The correlations were computed and also displayed with scattergrams, representative ones of which are included.

RESULTS

The extent to which the values of the approximations resembled their corresponding statistics is summarized in Table 2.

Standard Deviation:

For the standard deviation the largest difference in means was -.15. The approximations utilizing the sums of the upper and lower 1/6ths of the scores were notable in their correlation with the statistic (see also Figure 1) and in the low variability of the differences. The range divided by 5 provided an unbiased but more unpredictable estimate.

The error (i.e., the difference between approximation and statistic) was naturally related to number of testees for the standard deviation approximation based on the range, and as suggested in some statistics books, the constant could be changed according to sample size. For the first standard deviation approximation (SD 1/6), the largest error was from one of the smallest samples (N = 16), which accounted for the relationship of error with number of testees.

Tests with a large number of items tended to have the standard deviation slightly overestimated by the 1/6ths methods.

Lorraine K. Noval, Robert A. Pfeiffer, and James C. Ranney contributed to the analysis of the data.

Reliability:

The reliability estimates were compared with the results from the Kuder-Richardson Formula #20. The KR21 is of course an underestimate, but correlates highly with KR20. Saupe (1961) estimated mathematically that his approximations would tend to be closer to KR20 than would KR21. The present data tend to support his contention. The relationship between one of these Saupe approximations and KR20 is displayed in Figure 2. While two short-cut approximations (R23? and R16R) do more poorly overall than do the Saupe approximations using the variance, the substitution of a 1/6 standard deviation approximation into the traditional standard error formula (R1661 and R1761) resulted in quite dependable estimates.

None of the scattergrams of approximations with their statistics suggested non-linearity, although the results for the few tests with reliability coefficients less than .5 did suggest heteroscedasticity, as illustrated in Figure 2.

The relationship of average difficulty level to error for reliability estimates is noted in Figure 3; the shape for this estimate was the same as for the Saupe estimates and for the other reliability estimate where a 1/6th method standard deviation is used. Three of the four greatest underestimations occurred for the tests with the most extreme difficulty ($\bar{p}$ = .21, .83, .83). With the various reliability approximations, for the tests whose mean difficulty indices were above .4, many of the correlations between error and average difficulty level would have been in approximately the -.40s. While the Saupe approximations overestimated the reliability for some of the tests whose mean difficulty was in the .40s through .70s, the largest overestimate for any of the four formulae using the variance was only .106. In general the Saupe approximations tended to estimate conservatively.

For the Saupe reliability estimates the largest error was from one of the smallest samples (N = 31), as displayed in Figure 4, which helps explain the relationship of error with number of testees. For the reliability estimates, except for the KR21, the greatest variability of errors was for small samples.

For the shorter tests some of the reliability approximations tended to underestimate KR20.

Standard Error of Measurement:

The standard error of measurement approximation by Lord (1957, 1959) estimated the results of the standard formula with only slight errors. In addition, when his value of .43 was replaced by .4, the errors remained small. The correlations of the approximations with the exact statistic were slightly smaller than those reported by Lord for ETS tests, as may be noted visually from Figure 5.

Average difficulty level was related to performance of the Lord approximation (see Figure 6). The standard error was overestimated for the tests with $\bar{p}$ = .21 and .38, and also was overestimated for the easier tests. With the exception of the two most difficult tests, the Lord approximation became more conservative as the tests became easier. While the approximation was conservative for those two difficult tests, they are too few from which to generalize. (It should be noted that Lord limited his recommended interval to medium difficulty items). However, none of the errors reached 1.0 in absolute value. The scattergram for SE.4 was naturally very similar in shape to that for the Lord approximation, but the two difficult tests tempered the relationship considerably. For the longer tests, Lord's approximation slightly overestimated the standard error, as may be seen in Figure 7.

Summary:

Approximations for each of three statistics were computed and compared with the corresponding exact statistics. The one-sixth method for the standard deviation provided essentially the same results as did the exact statistic. The Mason-Odeh variation was slightly less biased, and indeed that bias would likely be reduced with N used in the exact statistic. Even the range divided by a constant provided an estimate which would often be suitable.

The Saupe approximations (using the variance) and the two with the standard formula incorporating a 1/6th approximation for the variance estimated KR20 with less bias than did KR21, although their relationship with KR20 was slightly lower. For tests of medium difficulty the errors were small, even when a one-sixth approximation was used to supply the variance.

The standard error was quite accurately estimated by using the number of items as a variable. Rounding Lord's constant to .4 resulted in a slightly less biased and less conservative estimate for these data, but to generalize to other sets of data would of course be hazardous. The results for the reliability and standard error approximations do imply that the Saupe and Lord approximations may be slightly short-cutted to make them even more suitable for non-computationally comfortable consumers and for scurrying statisticians.

DISCUSSION

From previous derivations, "rationalizations" and empirical data one would expect the approximations to perform well in practice. The results of the present study confirm and extend such expectations. For many practical applications approximations for the standard deviation, internal consistency reliability, and the standard error of measurement seem to provide sufficiently accurate estimates even where assumptions are not met.

One standard measure for reliability - KR21 - has been attacked recently by Cureton (1966) and Shoemaker (1969) as inappropriate for classroom tests. Saupe's (1961) approximations for estimating KR20 were supported by these data. The KR21 correlated highly with KR20 for this data, and thus a better approximation might be derived which would combine the lessened bias found for Saupe's approaches and the high correlation with KR20 found for KR21.

One limitation for some readers might be the definition of the standard deviation as the square root of the unbiased estimate of the variance, i.e., with N-1. Whether N or N-1 is more desirable depends to some extent on whether the estimate is conceptualized for a population value or a sample value. While the major conclusion about the suitability of approximations would not change, the relative superiority of some methods compared with some others might. The greatest effect is with the standard deviation approximations. The use of the more unbiased estimate favored slightly the one-sixth method with the N-1 (vs. N) in the denominator. Further work will need to be done to ascertain the extent of the effect. In this study mean sample size was 101, and therefore the effects of choice between N and N-1 would appear to have been minor.[1]

The choice of N vs. N-1 would likely have even less effect for comparing among reliability and standard error approximations. Using N-1 results in slightly larger KR20s and Saupe estimates, probably less than .01 for each when N=101, based on representative data. The (traditional) standard error would likely be little affected, both the standard deviation and the reliability coefficient are slightly larger with the use of N-1.

---

[1] With N=101, and $\sqrt{\frac{\Sigma x^2}{N}} = 5$, $\sqrt{\frac{\Sigma x^2}{N-1}} = 5.025$

The correlations of the differences with the skewness indicator, the absolute difference between the mean and the median, were low, with no particular pattern evident in the scattergrams. One reason for the low relationships could be the "quick and dirty" nature of the skewness indicator.

Further empirical studies using other sets of data should allow firmer conclusions about the quality of the approximations. Such studies of "robustness" should include measures of the third and fourth moments if possible.

CONCLUSION

Approximations for the standard deviation, internal consistency reliability, and the standard error of measurement were able to provide good estimates for this set of data. It appears reasonable to maintain that for practical situations approximations are adequate.

REFERENCES

CURETON, EDWARD E. Kuder-Richardson Reliabilities of Classroom Tests. Educational and Psychological Measurement. 1966, 26, 13-14.

DIEDERICH, P. B. Simplified Measurement Techniques for Teachers. In L.M. Huddleston (Ed.) Yearbook of the National Council on Measurements Used in Education. 1958, 24-27.

GARDNER, P. L. Test Length and the Standard Error of Measurement. Journal of Educational Measurement. 1970, 7, 271 273.

JENKINS, W. L. A Short-Cut Method for σ and r. Educational and Psychological Measurement. 1946, 6, 533-536.

LATHROP, R. L. A Quick-but Accurate-Approximation to the Standard Deviation of a Distribution. Journal of Exoerimental Education. 1961, 29, 319.321.

LORD, FREDERIC M. Do Tests of the Same Length Have the Same Standard Error of Measurement? Educational and Psychological Measurement. 1957, 17, 510-521.

LORD, FREDERIC M. Tests of the Same Length Do Have the Same Standard Error of Measurement. Educational and Psychological Measurement. 1959, 19, 233-239.

MASON, GEOFFREY P. and ODEH, ROBERT E. A Short-Cut Formula for Standard Deviation. Journal of Educational Measurement. 1968, 5, 319-320.

SABERS, DARRELL L. The Accuracy of Short-Cut Estimates for Standard Deviation of Raw Score Distributions on Teacher-Made Tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Minneapolis, March 1970.

SAUPE, JOE L. Some Useful Estimates of the Kuder-Richardson Formula Number 20 Reliability Coefficient. Educational and Psychological Measurement. 1961, 21, 63-71.

SHOEMAKER, DAVID M. Note on the Attenuating Effect of Zero-Variance Items on KR-20. Journal of Educational Measurement. 1969, 6, 255-256.

SWINEFORD, FRANCES Note on 'Tests of the Same Length Do Have the Same Standard Error of Measurement.' Educational and Psychological Measurement. 1959, 19, 241-242.

Table 1

THE STATISTICS AND THEIR APPROXIMATIONS

Label	Formula	Source
SD	$\frac{\Sigma x^2}{N-1}$	
SD1/6	$\frac{\Sigma \text{ upper } 1/6 - \Sigma \text{ lower } 1/6}{N/2}$	Jenkins-Diederich
SD6N1	$\frac{\Sigma \text{ upper } 1/6 - \Sigma \text{ lower } 1/6}{(N-1)/2}$	Mason and Odeh
SDR/5	range/5	
R	$KR20 = \frac{k}{k-1}\left[1 - \frac{\Sigma pq}{V_t}\right]$	Kuder-Richardson
R21	$KR21 = \frac{k}{k-1}\left[1 - \frac{k\bar{p}(1-\bar{p})}{V_t}\right]$	Kuder-Richardson
R19K	$1 = \frac{.19k}{V_t}$	Saupe
R18K1	$\frac{k}{k-1}\left[1 - \frac{.18k}{V_t}\right]$	Saupe
R17K1	$\frac{k}{k-1}\left[1 - \frac{.17k}{V_t}\right]$	Saupe
R16K	$1 - \frac{.16k}{V_t}$	Saupe
R23Q	$1 - \frac{2k}{23Q^2}$	Saupe
R16R	$1 - \frac{.16k}{(SDR/5)^2}$	McMorris
R1661	$1 - \frac{.16k}{(SD6N1)^2}$	McMorris

Table 1 (Continued)

Label	Formula	Source
R1761	$1 - \frac{.17k}{(SD6N1)^2}$	McMorris
SE	$SD\sqrt{1 - R}$	
SE3/7	$3/7\sqrt{k}$	Lord
SE.4	$.4\sqrt{k}$	
$\bar{p}$	$\frac{\bar{X}}{k}$	
Sk	$\lvert\bar{X} - \text{Mdn.}\rvert$	
N	Number of testees	
k	Number of items	

Symbol	Meaning
x	$X - \bar{X}$, deviation from the mean
Σ upper 1/6	sum of the upper 1/6th of the scores
p	proportion of testees answering an item correctly
q	$1 - p$, proportion of testees answering incorrectly
V_t	variance of scores on the total test
Q	semi-interquartile range, quartile deviation

Table 2

DIFFERENCES BETWEEN APPROXIMATIONS AND STATISTICS
DESCRIBED AND RELATED TO OTHER TEST CHARACTERISTICS

	Mean	S.D.	r with Statistic	Differences (Approx. - Stat.) Mean	Differences (Approx. - Stat.) S.D.	Correlation Coef's. for Differences with p	Sk	N	k
SD	5.84	2.30							
SD1/6	5.69	2.23	.99+	-.15	.19	-.06	-.20	.13	-.30
SD6N1	5.79	2.28	.99+	-.05	.17	-.16	-.16	-.09	-.12
SDR/5	5.85	2.41	.90	.02	1.07	.09	-.12	.61	.02
R	.70	.13							
R21	.60	.19	.98	-.10	.06	-.05	.15	-.00	.31
R19K	.62	.18	.94	-.08	.07	-.44	.14	.22	.29
R18K1	.65	.17	.92	-.05	.07	-.22	.14	.22	.09
R17K1	.67	.16	.91	-.03	.07	-.21	.11	.22	.03
R16K	.68	.16	.92	-.02	.06	-.19	.10	.24	.04
R23Q	.58	.20	.78	-.11	.13	-.04	.06	-.11	-.02
R16R	.65	.21	.74	-.04	.14	-.09	.00	.59	.10
R1661	.67	.16	.90	-.03	.07	-.22	.06	.19	.05
R1761	.65	.17	.90	-.05	.07	-.24	.09	.18	.10
SE	2.92	.75							
SE3/7	3.22	.78	.96	.30	.21	.25	.00	-.23	.33
SE.4	2.99	.74	.96	.08	.28	-.08	-.11	-.24	.19
p	.69	.09							
Sk	.57	.48							
N	101.02	105.41							
k	59.40	28.13							

FIGURE 1
SCATTERGRAM FOR THE STANDARD DEVIATION AND AN APPROXIMATION

X = SD Y = SDSM1
$\bar{X}$ = 5.84 $\bar{Y}$ = 5.79
$s(X)$ = 2.30 $s(Y)$ = 2.28
r = .997

Y																			
11.470																			
10.537																	4		4
10.037																2			2
9.537																			0
9.037												1			1				2
8.537													1	1					2
8.037												1							1
7.537											2	1							3
7.037										1	5	1							7
6.537										4	3								7
6.037									5	2									7
5.537								7	1										8
5.037							8												8
4.537						1	11	1											13
4.037						3													3
3.537					3														3
3.037				4															4
2.537			1	3															4
2.037		4																	4
1.710	3																		3
	2.086	3.086	4.086	5.086	6.086	7.086	8.086	9.086	10.086	11.086									
	1.660	2.586	3.586	4.586	5.586	6.586	7.586	8.586	9.586	10.586									
	3	5	3	4	3	4	11	9	7	6	7	8	3	3	1	1	1	2	4	85

FIGURE 2

SCATTERGRAM FOR KUDER-RICHARDSON 20 AND AN APPROXIMATION

$X = R$ $\quad Y = R16K$

$\bar{X} = .70$ $\quad \bar{Y} = .68$

$s(X) = .13$ $\quad s(Y) = .16$

$r = .92$

Y																	
.923																1	1
.900													5	1			6
.850										1	3	4	1				9
.800								1	5	5							11
.750						1	6	10	1								18
.700						3	4	4									11
.650				2	1	2	4						1				10
.600			1		3												4
.550			1	3	2	1											7
.500			2														2
.450		1															1
.400				1	1												2
.350																	0
.300																	0
.250	1				1												2
.200				1													1

.250	.271	.321	.371	.421	.471	.521	.571	.621	.671	.721	.771	.821	.871	.921	.940	
1	1	0	2	2	4	5	6	7	15	20	9	9	3	1		85

FIGURE 3

SCATTERGRAM FOR AVERAGE DIFFICULTY VS. A RELIABILITY DIFFERENCE

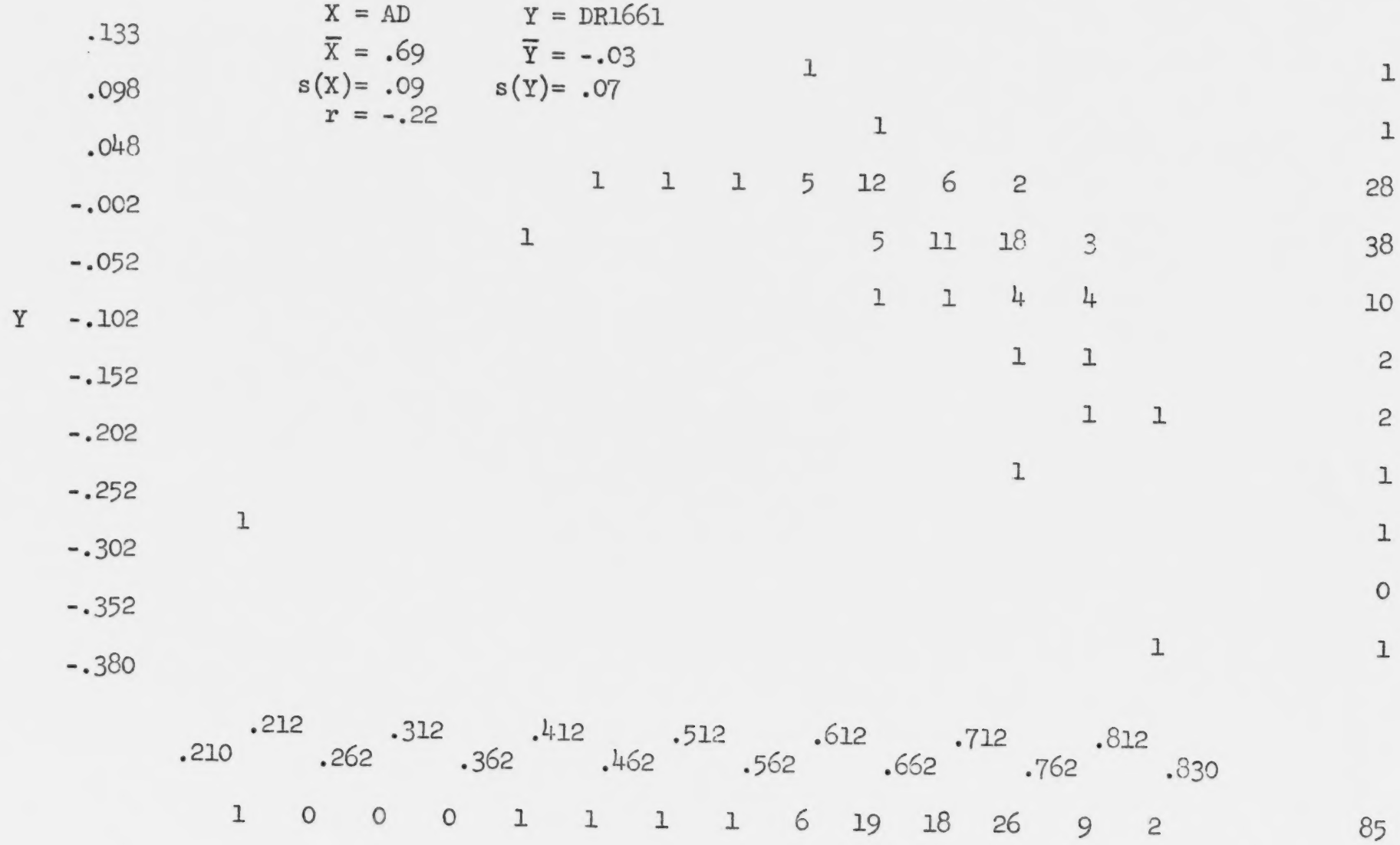

FIGURE 4
SCATTERGRAM FOR NUMBER OF TESTEES VS. A RELIABILITY DIFFERENCE

X = N $\bar{X}$ = 101.02 s(X) = 105.41 r = .24
Y = DR16K $\bar{Y}$ = -.02 s(Y) = .06

Y																
.099	1															1
.091														1		1
.066		1			1			1			1					4
.041	1	2	2				4	2			1			1		13
.016		7	4	1	1	2		4	1	1	1					22
-.009	2	6	3	4	2	2			1							20
-.034	1	3	4	3	3		1	1								13
-.059	1		2							1						4
-.084		2												1		3
-.109																0
-.134		1														1
-.159																0
-.184																0
-.209					1											1
-.234																0
-.259	1	1														2
-.335																
	16.00	31.02	51.02	71.02	91.02	111.02	131.02	151.02	171.02	191.02	211.02	231.02	251.02	271.02	291.02	311.02
	7	23	15	8	5	4	5	8	2	2	3	0	0	0	3	85

FIGURE 5
SCATTERGRAM FOR THE STANDARD ERROR OF MEASUREMENT AND AN APPROXIMATION

X - SE Y - SE3/7
$\bar{X}$ - 2.92 $\bar{Y}$ - 3.22
s_X - .75 s_Y - .78
r - .96

Y	1.400	1.540	1.790	2.040	2.290	2.540	2.790	3.040	3.290	3.540	3.790	4.040	4.290	4.540	4.600	
5.100														1		1
5.095																0
4.845									1		2	1				4
4.595																0
4.345							2		1	4						7
4.095							5	1	1							7
3.845						2	1	2								5
3.595					1		3	4	1							9
3.345				2	6	13	6									27
3.095						2										2
2.845				3	2	1										6
2.595					1	3										4
2.345					2	3										5
2.095			1	4	1											6
1.845		1														1
1.595		1														1
1.400																
	3	4	4	6	6	8	21	10	9	3	3	4	2	2		85

FIGURE 6

SCATTERGRAM FOR AVERAGE DIFFICULTY VS. A STANDARD ERROR OF MEASUREMENT DIFFERENCE

X = AD $\bar{X}$ = .69 s(X)= .09 r = .25

Y = DSE3/7 $\bar{Y}$ = .30 s(Y)= .21

Y	.210	.262	.312	.362	.412	.462	.512	.562	.612	.662	.712	.762	.812	.830	
.900														1	1
.855													2	1	3
.755													1	1	3
.655												1			1
.555											1	4	3	2	10
.455										4	8	3			15
.355									3	4	5	1			13
.255									7	8	5				20
.155								2	8		1				11
.055							1	1	4	1					7
-.045						1									1
-.100															
	1	0	0	0	1	1	1	1	6	19	18	26	9	2	85

FIGURE 7

SCATTERGRAM FOR NUMBER OF ITEMS VS. A STANDARD ERROR OF MEASUREMENT DIFFERENCE

X = K Y = DSE3/7
$\bar{X}$ = 59.40 $\bar{Y}$ = .30
s(X) = 28.13 s(Y) = .21
r = .33

Y	10.000–14.400	14.400–24.400	24.400–34.400	34.400–44.400	44.400–54.400	54.400–64.400	64.400–74.400	74.400–84.400	84.400–94.400	94.400–104.400	104.400–114.400	114.400–124.400	124.400–134.400	134.400–140.000	
.900											1			1	
.855									2		1			3	
.755				2			1							3	
.655						1								1	
.555		1	1		4	1		2					1	10	
.455			2	1	2	5	2	3						15	
.355			1	2	2	5		1	1		1			13	
.255		6	1		4	6			2	1				20	
.155			1	2	1	1		3		2		1		11	
.055		1		1	1	2	1	1						7	
-.045					1									1	
-.100															
	1	8	8	6	17	18	4	10	0	7	1	4	0	1	85